Memory

What is memory?

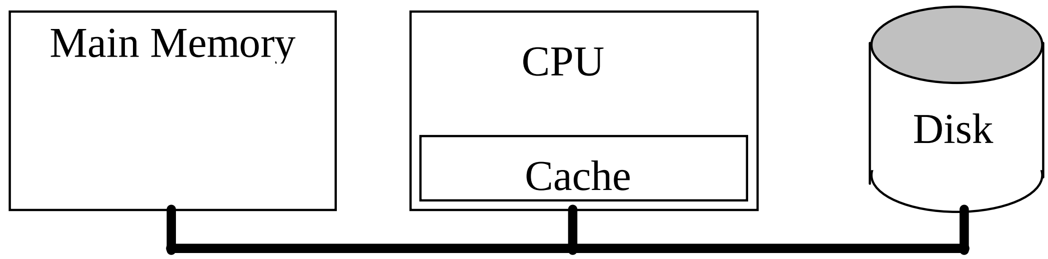
Memory is defined as volatile space used to hold execution program code and data. In this section we will discuss main memory (DRAM), cache memory (SRAM), the latest advances in memory and the terms often associated with memory.

A computer will use two types of memory, main memory and cache. Main memory is often referred to as Dynamic Random Access Memory (DRAM). DRAM is used as a vast memory pool to hold the bulk of the applications and data. It is inexpensive and relatively slow. Unlike DRAM, Cache is a relatively small amount of memory which is costly and fast. Cache is referred to as Static Random Access Memory (SRAM). It speeds up processing by being placed near the processor. The processor can then operate from cache instead of going to main memory for each request. Cache is used by the processor to hold the current application code and data which is being used directly by the processor. The various levels of cache are referred to as 1^{st}, 2^{nd}, and 3^{rd} level cache.

The process of moving code and data to cache has improved dramatically. In early designs, a processor would request data, the disk would return data to the processor, the processor would put the data into memory and would extract small portions of the data to cache. Today, a Direct Memory Access chip assists with many memory requests allowing the processor to do other work. The processor is alerted when the process is complete.

The processor takes data from disk and places it into main memory. Data that is being processed immediately is moved to cache.

CPU, Main Memory, Cache, and Disk



How Memory Is Used

Let us use an analogy to help explain how memory is used. Main memory, 2nd level cache, and 1st level cache could be compared to a supermarket, a refrigerator, and a dinner table.

Most of us buy our food at the supermarkets which are large repositories of items we need. We then take these items (food) home and deposit them in our refrigerator which is closer to where we need it. As a final step we remove the items from the refrigerator to the dinner table where we can consume (process) the food.

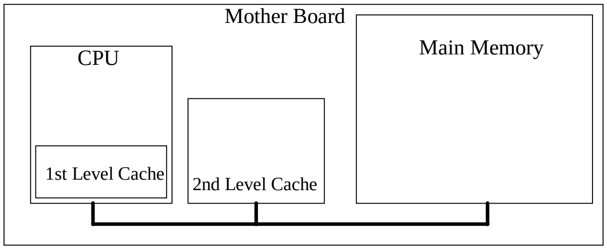
Just as you would buy the bulk of your food at the supermarket, a processor retrieves items we need (data) and stores it in main memory. Main memory, like the supermarket, are large areas where you have a variety of items from which to choose. And like the refrigerator, 2nd level cache acts as a local storage area to store code and data that will most likely be utilized in the near future.

To get to that final step we have to move the consumable to the processing point, the dinner table, or as in our computer analogy, 1st level cache. 1st level cache is a small speedy storage spot used by the processor to process the most recently used instructions. Eventually, all code and data being used moves from main memory down the chain to 1st level cache and finally is processed by the CPU. The key to memory management is to have a processor find most of its information in 1st or 2nd level cache. This reduces requests to main memory.

In the beginning the processor was separated from the 1st level cache which was on the motherboard. In subsequent generations, the 1st level cache was attached to the processor, and the 2nd level cache was on the motherboard. The 2nd level cache was still closer to the processor than main memory and operated much faster than main memory. Current trends are to connect the 1st and 2nd levels of cache directly to the processors, and if needed, a 3rd level of cache can be added on the motherboard (usually to tie between clusters of processors).

1. A processor requests data and checks 1st level cache. It is often found and processed.
2. If not found, the processor then checks 2nd level cache.
3. If not found it then checks main memory.
4. If still not found the processor makes a request to the operating system to retrieve the data from disk.
5. Once the data moves up the chain into 1st level cache it is processed.

A CPUs Request for Code or Data



Memory Speeds, Feeds, and Terms

In an attempt to understand the speed at which memory works, let us consider disk speeds and memory speeds. Disk speeds are measured in milliseconds (ms). Memory is measured in nanoseconds. A nanosecond is one *billionth* of a second. Incredibly, memory speeds can be as much as 1000 times faster than disk speeds.

DRAM

Main memory is referred to as Dynamic Random Access Memory (DRAM). DRAM is relatively inexpensive and its speeds are measured in nanoseconds. A typical speed for DRAM is about 60 nanoseconds (60ns). One of the reasons that DRAM is slower is because it needs to be refreshed.

SRAM

Static Random Access Memory (SRAM) is the memory used in cache. This memory does not need to be refreshed and is much faster than DRAM, although the speed comes at a price. Faster cache is more expensive, which is why you will see systems with large amounts of DRAM and small amounts of cache.

EPROM

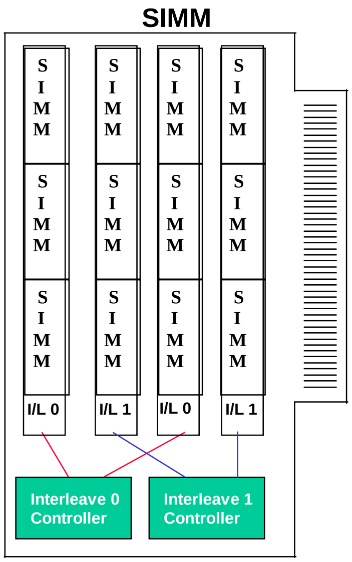
Erasable Programmable Read Only Memory. A type of read-only memory (ROM) where the data pattern may be erased to allow a new pattern. EPROMs are usually erased by ultraviolet light and recorded by a higher than normal voltage programming signal.

Flash

A type of EPROM that can be re-programmed by the computer or peripheral to which it is connected.

SIMM

Single In-line Memory Module. SIMMs modules are placed on a DRAM memory board in order to add memory. Think of SIMMs as checkers and a memory board as a checker board. You have the ability to add memory by placing additional checkers on the board until the entire board is filled. SIMMs also come in different density sizes. If you use checkers that are twice as thick you will eventually double your memory capacity.

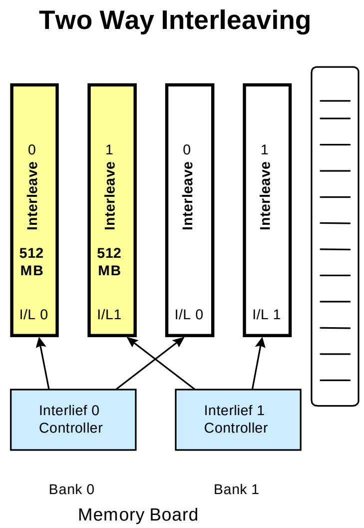


Interleaving

In interleaved memory, memory is divided into a set of banks. An interleaved memory with *X* number of banks is said to be *X*-way interleaved. Think of interleaving as parallel processing for memory.

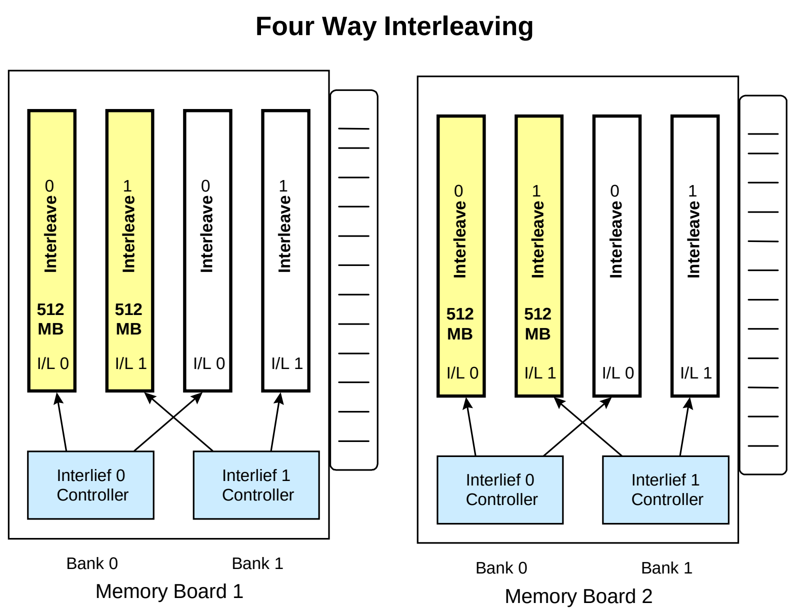
Non-interleaved memory divides its consecutive addresses into contiguous blocks. In interleaved memory, consecutive addresses reside in different banks. For example, suppose there are four banks, each containing 512 bytes. The block oriented scheme would assign addresses 0 through 511 to the first bank, 512 through 1023 to the second bank, and so on until all the memory was used. The interleaved scheme would assign addresses 0,4,8,12…to the first bank, and 1,5,9,13….to the second bank until all the memory was used.

However the memory is split up among the banks, as long as requests are sent to two different banks, they can be handled simultaneously. The result is improved bandwidth.



This memory board has two banks for SIMM or DIMM memory. Each bank is subdivided into two interleaves, 0 and 1. The memory board has two controllers, one for each interleave.

Each bank is independent of the other. By populating the memory board with equal memory sizes in interleave 0 and interleave 1 you have two way interleaving.



By populating two memory boards with equal amounts of memory in the first bank in interleaves 0 and 1 you have four way interleaving. The four interleaves must be equally populated and their memory addresses are spread in order.

As Interleaving increases, performance increases. You can gain about 12-15% in performance with four-way interleaving. If all of the banks in the picture were populated, the memory would be said to be 8-way interleaved.

DIMM

DIMMs stands for Dual In-line Memory Module. A DIMM utilizes both sides of the memory board giving you twice the memory on the same real estate. DIMMs also come in different densities. DIMMs utilize a 128-bit data path through interleaving while SIMMs utilize 64-bits.

Extended Data Out Dynamic Random Access Memory (EDO DRAM)

EDO DRAM is one of a series of recent innovations in DRAM chip technology. Only certain computer systems are designed with memory controllers setup to support it. EDO memory allows a CPU to access memory 10% to 15% faster than comparable fast-page mode chips. EDO memory allows the data outputs to take advantage of pipelined systems that overlap accesses where the next cycle is started before the data from the last cycle is removed from the bus.

Synchronous Dynamic Random Access Memory (SDRAM)

SDRAM is a form of DRAM which adds a separate clock signal to synchronize signal input and output on a memory chip. The clock is coordinated with the CPU clock so the timing of the memory chips and the timing of the CPU are synchronized. Synchronous DRAM saves time in executing commands and transmitting data, thereby increasing the overall performance of the computer.

Rambus Dynamic Random Access Memory (RDRAM)

RDRAM is a form of DRAM that is used mainly for video accelerators. It offers sustained transfer rates of 5 times that of normal DRAM. Although it cannot be used as a direct replacement for existing memory, it is likely that it will replace DRAM and SDRAM as the main memory system in personal computers as the bus speeds required by these machines increase. SDRAM can operate up to 100 MHz, but RDRAM has been demonstrated at 600 MHz. This memory is only 8-bits wide, so the bandwidth would increase enormously if it were used in parallel to give 32 or 64-bit memory.

Keys Points To Remember

- The two most common forms of memory are Dynamic Random Access Memory (DRAM) and Static Random Access Memory (SRAM).
- DRAM is used in main memory. DRAM is slower and less expensive than SRAM.
- Cache or SRAM is used so that the processor does not have access main memory unless needed. Cache is fast and expensive.
- SIMMs stands for Single In-line Memory Module. This is the most popular way to populate memory boards.
- DIMMs stands for Dual In-line Memory Module. This is replacing SIMMs because you get twice the memory for the same real estate.
- Interleaving is a process of placing equal amounts of memory in a memory board's bank 0 and bank 1. The system can use two controllers to parallel process the memory.